Technical Report 423

LEVEL II

12

ADAPTIVE MENTAL TESTING: THE STATE OF THE ART

James R. McBride

PERSONNEL UTILIZATION TECHNICAL AREA

U. S. Army
Research Institute for the Behavioral and Social Sciences

November 1979

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

FRANKLIN A. HART
Colonel, US Army
Commander

NOTICES

DISTRIBUTION Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TP, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report 423	2. GOVT ACCESSION NO. AD-A088 000
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): ADAPTIVE MENTAL TESTING: THE STATE OF THE ART,	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER: --	
7. AUTHOR(s): James R. McBride	8. CONTRACT OR GRANT NUMBER(s): --
9. PERFORMING ORGANIZATION NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q162717A766
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Deputy Chief of Staff for Personnel, Washington, D.C. 20310	12. REPORT DATE: November 1979 13. NUMBER OF PAGES: 39
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Adaptive tests	Computer testing	Aptitude tests
Mental testing	Psychometrics	
Testing	Tailored tests	
Mental measurement	Latent trait theory	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper defines adaptive mental testing in relation to conventional mental testing, outlines the major research issues in adaptive mental testing, and reviews the state of the art for each of the research issues. The research issues are: (1) psychometric theory; (2) design of adaptive tests; (3) scoring adaptive tests; (4) the testing medium; (5) item pool development; and (6) advances in measurement technology.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Technical Report 423

ADAPTIVE MENTAL TESTING: THE STATE OF THE ART

James R. McBride

Submitted by:
M. A. Fischl, Acting Chief
PERSONNEL UTILIZATION TECHNICAL AREA

Approved by:

E. Ralph Dusek
PERSONNEL AND TRAINING
RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

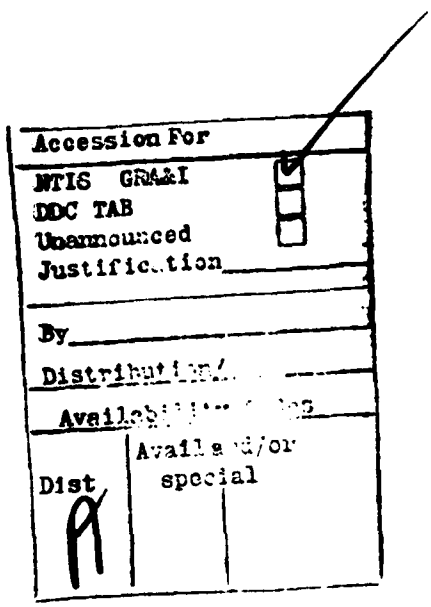
Office, Deputy Chief of Staff for Personnel
Department of the Army

November 1979

Army Project Number
2Q162717A766

Manpower Systems
Technology

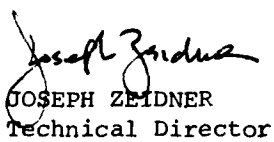
ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.



FOREWORD

The Personnel Utilization Technical Area of the Army Research Institute for the Behavioral and Social Sciences (ARI) is concerned with developing more effective techniques for measuring people's abilities, to aid in Army job assignment. An emerging technology which offers considerable promise in this area is computer-based adaptive mental testing. This report was prepared under Army Project 2Q162717A766, Manpower Systems Technology, to identify technology gaps and deficiencies and to summarize new trends in the state of the art of mental testing.

The report was prepared while the author was a staff member of ARI. He is presently on the staff of the Naval Personnel Research & Development Center, San Diego, Calif.



JOSEPH ZEIDNER
Technical Director

ADAPTIVE MENTAL TESTING: THE STATE OF THE ART

BRIEF

Requirement:

To identify technology gaps and deficiencies and to summarize new trends in the state of the art of mental testing.

Procedure:

Adaptive mental testing is defined in relation to conventional mental testing. The state of the art is assessed for each of six research issues in adaptive mental testing: (1) psychometric theory; (2) design of adaptive tests; (3) scoring adaptive tests; (4) the testing medium; (5) item pool development; and (6) advances in measurement technology.

Findings:

Specific research requirements are identified for each research issue in adaptive mental testing. Discussion of these requirements is also provided.

Utilization of Findings:

This research forms a basis for designing a research and development program for application of adaptive mental testing technology to military applicant selection and job assignment.

ADAPTIVE MENTAL TESTING: THE STATE OF THE ART

CONTENTS

LIST OF TABLES

ADAPTIVE MENTAL TESTING: THE STATE OF THE ART

INTRODUCTION

The measurement of psychological traits is usually accomplished by observing the responses of examinees to selected test items. For some traits, notably the ability/aptitude traits assessed during personnel selection and classification, all examinees are required to answer a common set of items, and the test score is a linear composite of the dichotomous item scores. This test score is used as an index of individual differences to differentiate among the persons tested.

It has long been known that administering the same test items to all persons--as is done in conventional group tests--provides less than optimal discriminability, and that the ability to differentiate accurately among persons of varying trait status could be enhanced by individually tailoring the test items to the status of the examinee. In ability measurement terms, this connotes dynamically tailoring test item difficulty to the ability level of the individual. A test that proceeds in this fashion is called an adaptive, or tailored, test (Weiss & Betz, 1973; Wood, 1973). Adaptive tests have striking psychometric advantages over conventional tests under certain circumstances, and they have aroused considerable interest among test theoreticians.

The development of adaptive testing has been motivated largely by recognizing that conventional group ability tests do not measure individual differences with equal precision at all levels of ability; this is because accuracy and precision of measurement are in part a function of the appropriateness of test item difficulty to the ability of the individual being measured.

To measure with high precision at all levels of ability requires tailoring the test--by either item difficulty or test length, or both--to the individual. Since ability is unknown at the outset of testing, the tailoring process must be done during the test; hence the requirement for adaptive ability testing. This is done by choosing test items sequentially, during the test, to adapt the test to the examinee's ability as shown by responses to earlier test items. This can be done by a human examiner, using paper-and-pencil tests with special instructions, or by means of a mechanical testing device. The device most commonly used is an interactive computer terminal.

The motivation for adaptive testing is that it should permit measuring ability with higher and more equal precision throughout a wide ability range than can conventional group tests in which all persons answer the same test items. In terms of classical psychometric indices, improved measurement in that sense should be accompanied by corresponding improvements in reliability and in external validity. In addition to the psychometric benefits, there are potential psychological benefits to examinees

in the reduction of frustration or boredom resulting from adapting test difficulty to the individual.

The rationale behind adaptive testing has existed for years. The Stanford-Binet intelligence test is an adaptive test, administered personally by a skilled examiner. Mass testing using adaptive methods would make such personal administration impractical, however. The development of adaptive testing awaited the availability of testing media that would permit widespread use of adaptive tests on a fairly large scale. A number of problems--psychometric and technological--had to be solved before adaptive testing could be practical on a large scale. This paper contains a review of some of those problems, and a summary of the state of the art in research addressing them.

BACKGROUND

Conventional Test Design

Conventional group administrable tests of psychological variables, such as mental abilities, involve administering of a common set of items to all examinees. The total score on such tests, usually the number correct or some transformation thereof, is used to index individual differences on the variable being measured. This procedure has been sanctified by longstanding practice and by empirical usefulness, but it has disadvantages as a measurement technique.

To construct a conventional test, the test designer chooses some subset of items from a larger pool of available items known to measure the variable of interest. Since the items in the pool typically vary in their psychometric properties--particularly in their difficulty--the test designer must decide what configuration of these item psychometric properties best suits the test's purpose. There are two extreme rationales to guide that decision. One rationale is to choose items that are highly homogeneous in item difficulty. A test so constructed, called a "peaked" test, will discriminate very effectively over a narrow range of the variable, but will discriminate poorly outside that range. The purpose of a peaked test design is to make fine discriminations in the vicinity of a cutting point; e.g., to categorize examinees into "go" and "no-go" groups for selection purposes.

At the opposite extreme is the "uniform" test, constructed of items that are heterogeneous in difficulty, with item difficulty parameters spread over a wide range. A uniform test will discriminate with more or less equivalent precision over a wide range of the variable, but (other things being equal) the level of precision will be substantially lower than that of the peaked test at the latter's best point. The purpose of a uniform test is to measure with equal precision throughout a wide range of the trait; e.g., to obtain information on which to aid assignment decisions to jobs requiring varying amounts of the tested ability.

In constructing a conventional test of given length, the test designer must choose between high precision over a very narrow range, and low to moderate precision over a wide range. A test cannot have both high precision and wide range unless the test is very long or the item difficulty is tailored to each examinee's level on the underlying variable. The use of long tests is often impractical. The alternative--tailoring test difficulty to each examinee--represents a striking departure from conventional group testing practice.

Adaptive Test Design

In an adaptive test, the test administrator chooses test items sequentially during the test, in such a way as to adapt test difficulty to examinee ability as shown during testing. An effectively designed adaptive test can resolve the dilemma inherent in conventional test design. By tailoring tests to individuals, the adaptive test can approximately achieve the high point precision of a peaked test and can extend that high level of precision over the wide range of a uniform test. As a result, a well-constructed adaptive test should be more broadly applicable than a conventional test of comparable item quality and test length, since its precision characteristics make it useful for classification about one or many cutting points, as well as for measurement over a wide range.

It is important to understand how an adaptive test can achieve psychometric advantages over conventional tests. It can be shown that measurement error is a function of the disparity between item difficulty and personal ability, as well as the discriminating power of the test items and their susceptibility to guessing. Since a peaked test concentrates item difficulty at a single ability level, measurement error should be smallest at that critical level, and increasingly larger at ability levels deviant from the critical point. In the case of a uniform test, item difficulty is spread over a wide range; consequently, measurement error tends to be low to moderate and fairly constant over a correspondingly wide range.

What is desirable, of course, is to achieve small measurement error over a wide range of the trait scale. This can be done only by administering items of appropriate difficulty at every ability level of interest. The rationale of adaptive testing is to do this more efficiently (i.e., in fewer items) than can be done by conventional means. This implies individualized choice of test items for each examinee. Administratively, this can be accomplished (a) by individual testing by skilled examiners, (b) by specially designed group-administered paper-and-pencil adaptive tests with rather complex instructions,[1] or

---

[1] An example of this kind of test is the flexilevel test devised by Lord (1971a).

(c) by automated testing using a computer or a specialized stimulus programmer to choose and administer test items. Research in adaptive testing has emphasized computer-controlled test administration.

Early research pertinent to adaptive testing was reviewed by Weiss and Betz (1973), and by Wood (1973). Subsequent research has been reviewed by this writer (McBride, 1976a). Research in adaptive testing has progressed from exploratory studies of item branching tests (e.g., Seeley, Morton, & Anderson, 1962), through the explication of a novel test theory applicable to tailored tests (e.g., Lord, 1970, 1974a), to the verge of operational implementation of a large-scale adaptive testing system for personnel selection (Urry, 1977b).

From a psychometric viewpoint, adaptive tests are attractive for a number of reasons. Adaptive tests represent a breakthrough in the technology of psychological measurement, because they can yield more precise measurement over a wider range with substantially fewer items than can conventional tests. In other words, adaptive tests can achieve higher validity of measurement than comparable conventional tests in a given test length; or, they can attain a given level of validity in substantially fewer items than a comparable conventional test (Urry, 1974).

Other aspects of adaptive tests also make them attractive, particularly if they are computer-administered. Tailoring test difficulty to examinee ability may reduce error variance caused by examinee frustration, boredom, or test anxiety (Weiss, 1974), as well as by guessing. Computer administration and scoring can reduce human error in marking answers, scoring the tests, and recording the results. Test compromise can be reduced substantially, by eliminating test booklets (thus negating theft) and by individualizing test construction (thereby thwarting the use of cheating devices). Printing, storage, and handling of test booklets and answer sheets can be eliminated, saving costs.

The psychometric and practical potential of adaptive testing makes it worthy of research and development in the military manpower setting, with the goal of eventual implementation of an automated system for test administration and scoring, and personnel selection, classification, and job-choice counseling. Some of the relevant research has already been done and has been reviewed as cited above. One outcome of the completed research has been the crystallization of a number of research issues that need to be resolved before deciding whether to implement an adaptive testing system. The purpose of this report is to present some of those issues and to evaluate the state of the art with respect to their resolution.

RESEARCH ISSUES

Psychometric Theory

Early adaptive testing research showed that traditional test theory was an inadequate basis for the construction and scoring of adaptive tests (e.g., Bayroff & Seeley, 1967). This was due to requirements for item parameters that were invariant with respect to examinee group, and means of scoring tests in which different examinees answered sets of items that differed in difficulty, number, and other respects as well. One resolution of this issue was provided by the earlier development of item response theory (Rasch, 1960; Lord, 1952, 1970, 1974a; Birnbaum, 1968) that provided the needed invariance properties for item parameters and test scoring capabilities.

Subsequent approaches to adaptive testing were developed that did not depend on the rather strong assumptions of item response theory. Kalisch (1974) and Cliff (1976) both presented theory and methods for adaptive testing that are not based on the stochastic response models of item response theory. Other psychometric bases appropriate for use in adaptive testing may be forthcoming. Clearly, one research issue to be addressed is the adequacy of the psychometric foundation of any proposed approach to the implementation of adaptive testing.

Item Response Models

Most adaptive testing research since 1968 has used item response theory (item characteristic curve, or latent trait, theory) as a psychometric basis. Within item response theory, several competing response models for dichotomously scored items have been proposed. These models differ in mathematical form and in the number of parameters needed to account for item response behavior. Some of these models include the one-parameter Rasch logistic model (e.g., Wright & Douglas, 1975); the two-parameter normal ogive model (Lord & Novick, 1968); and the three-parameter logistic ogive model (Birnbaum, 1968). These models differ in mathematical complexity and in the procedures required to implement them in practice. If adaptive testing research is to be based on item response theory, a consequent research issue is to choose from among the available response models the one best for the purpose. The basis for such a choice should include consideration of the appropriateness of the competing models, their robustness under violations of relevant assumptions, and the difficulty and expense of implementing them.

Design of Adaptive Tests

Strategies for Adaptive Testing

Adaptive testing by definition involves sequential selection of the test items to be answered by each examinee. Numerous methods for

sequentially choosing items have been proposed. These methods, called "strategies" for adaptive testing, were reviewed by Weiss (1974). Since then, several new ones have come forth (e.g., Cliff, 1976; Kalisch, 1974; McBridge, 1976b).

These strategies vary along a number of dimensions, including mathematical elegance, item selection algorithms, scoring methods, and others. There is a clear need for research to compare the various strategies on their psychometric and practical merits to provide the data needed to guide a choice among strategies.

Test Length

Any mental test has some criterion for test termination--a rule for stopping. Usually, a power test terminates when the examinee has answered all the items (although a time limit may be imposed for administrative convenience). Some adaptive testing strategies also use fixed test length as a stopping rule: Terminate testing when the examinee has answered some fixed number of items. Other strategies for adaptive testing, however, allow test length to vary from one examinee to another by basing the termination decision on some criterion other than test length. For example, testing may be terminated when a ceiling level of difficulty has been identified (e.g., Weiss' (1973) stratified adaptive strategy), or when a prespecified degree of measurement precision has apparently been attained (e.g., Urry, 1974; Samejima, 1977).

The research issue here concerns the relative merits of fixed length versus variable length adaptive tests. Is one alternative generally preferable over the other or preferable for some testing purposes but not for others? The notion of variable length tests has some intuitive appeal. Research is required to verify whether variable length tests have psychometric and practical merit.

Test Entry Level

Another aspect of the design of adaptive tests is test entry level--the difficulty level of the first item(s) the examinee must answer. In some cases there may be reliable information available prior to testing that would justify the use of different starting points for different examinees. For example, in a multitest battery, some subtests are substantially intercorrelated; an examinee's score on an early subtest may provide useful data for choosing entry level on a subsequent subtest.

The use of differential entry levels may permit us to improve measurement accuracy or to achieve a given level of measurement accuracy in even fewer items than an adaptive test that uses a fixed entry level. Research is needed to determine if these potential advantages of differential test entry level can be achieved.

Scoring Adaptive Tests

Because an adaptive test is fundamentally different from a conventional test in which everyone answers the same questions, it follows that conventional test scoring methods may not be applicable to adaptive tests. That is, it may make little sense to score an adaptive test by weighting and summing the dichotomous item scores. If so, alternative scoring methods are needed, which gives rise to yet another research issue: What means of scoring adaptive tests are available, and which are "best" in some important sense?

A related issue is the comparability of scores on adaptive tests with more familiar scores on standardized conventional tests. Are appropriate score equating methods available for transforming adaptive test scores into the metric of raw or converted scores of established conventional measures having the same variables?

The Testing Medium

Conventional ability tests are typically administered via paper and pencil, and constructed of multiple-choice items. Adaptive tests using the same item types may be administered individually (a) by a skilled examiner, (b) at an automated testing terminal, perhaps controlled by a computer; or (c) by means of specially constructed paper-and-pencil tests.

Individual testing by skilled examiners is impractical for large-scale use. Thus, only automated testing terminals and specially designed paper-and-pencil tests merit serious consideration as potential media for adaptive testing on a large scale. Whether paper-and-pencil adaptive testing is even feasible is problematic because of the requirement for sequential item selection. Another research issue, then, concerns the feasibility of group administration of paper-and-pencil adaptive tests.

The feasibility of automated test administration is not in question, since the presentation of test items and the recording and processing of an examinee's responses can be done using modern computers with interactive visual display terminals, such as teletype, cathode ray tube (CRT), or plasma tube (PLATO) terminals.

Nevertheless, computers and computer terminals are presently relatively expensive compared to traditional printed test booklets and answer sheets. It may be preferable to base automated adaptive tests on devices that are somewhat less sophisticated and less costly than full-scale computer systems. Still another research issue surfaces here: What alternative devices/systems may be used for automated adaptive testing, and what are the advantages and disadvantages of each?

Item Pool Development

Selecting the items to constitute an adaptive testing item pool is a somewhat larger undertaking than choosing items for a conventional test. The psychometric criteria for item selection and for pool construction are more rigorous than those for conventional test design, and the item pool must be substantially larger than the length of any individualized test drawn from it. Since the degree to which an adaptive test realizes its potential may be limited by the size and quality of its item pool, it is imperative that research defines the necessary or desirable characteristics of item pools for adaptive testing and provides practical prescriptions for item pool development.

Advances in Measurement Methodology

Adaptive administration of traditional dichotomously scored test items promises a significant gain in the psychometric efficiency of measurement. Since adaptive testing research has stressed the use of computer terminals for test administration, we should exploit the unique capabilities of computers to control test situations that are vastly different from the relatively simple tasks that comprise paper-and-pencil tests. New approaches to ability measurement may arise from the conjunction of adaptive test design and computerized test administration, and thus a number of research issues may arise. These issues could include the following: How can the expanded stimulus and response modes made possible by computer administration be exploited to improve the measurement of traditional ability variables? What new variables can be identified and measured using the computer's unique capabilities? Are scaling techniques available that are appropriate for those new measures? How does the utility of new measurement methods compare with that of traditional testing?

THE STATE OF THE ART

The problems originally hindering the development and implementation of adaptive testing were (a) psychometric and (b) practical. The psychometric problems concerning adaptive tests included the inappropriateness of classical test theory, the lack of prescriptions for their design, the need for methods of scoring, and the need for assessing the measurement properties. The practical problems included the need to develop new media for administering adaptive tests and the difficulty of assembling the large pools of test items demanded. Each of these problems will be discussed below, followed by a brief exposition of the state of the art relevant to solution of specific problems.

Psychometric Theory

Discussion

Traditional, or classical, test theory is inadequate to deal with some of the psychometric problems posed by adaptive tests. The problem in classical test theory was to order persons with respect to an individual differences variable on the basis of their number correct or proportion correct on common or equivalent tests. The observed score was assumed to differ from the "true score" by a random variable that was uncorrelated with true score. In adaptive testing, different persons respond to sets of test items that are in no sense equivalent across persons. These individualized tests may differ in difficulty, length, and the discriminating powers of their items. Obviously, the number or proportion of correct scores is generally an inappropriate index of individual differences; additionally, measurement error cannot be assumed to be independent of the variable being measured. A test theory was needed that could accommodate the special requirements of adaptive tests.

Several solutions to this problem might be forthcoming. A class of solutions currently exists, in the body of latent trait mental test theories, or item response theory. These "theories" are actually statistical formulations that account for test item responses in terms of the respondent's location on a scale of the attribute being measured by the item. The best developed formulations to date deal with dichotomous item responses as functions of a unidimensional attribute variable.

In the language of ability and achievement testing, latent trait methods treat the probability of a correct response to a test item as a monotonic increasing function of the relevant underlying ability. When a scale for the ability is established, the latent trait methods provide mathematical models relating response probability to scale position. These models are item trace lines, or item characteristic curves (i.c.c.).

Once a scaling of the attribute has been accomplished and all the item characteristic functions are known, the location of an individual on the attribute continuum can be estimated statistically from the dichotomously scored responses to any subset of the test items. Such an estimate is a kind of "test score"; the advantage of using latent trait methods for scoring is that all scores are expressed in the same metric, regardless of the length or item composition of the test. Thus, within the limits of the method, automatic equating of different tests can be effected merely by using latent trait methods for scoring the tests. This feature makes latent trait test theory an especially appropriate basis for adaptive testing.

The prevailing trend in application of latent trait methods has been to scale the measured attribute in such a way that all item characteristic curves have the same functional form, differing from item to item only in the parameters of the item characteristic functions.

Thus, once the general functional form has been established, each test item can be completely characterized and differentiated from other test items by the parameter(s) of its i.c.c. For attributes such as ability and achievement variables, where item trace lines should be monotonic in form, several similar response models have been developed in detail. These include a one-parameter logistic ogive model due to Rasch (1960), of which Wright (1968; Wright & Panchapakesan, 1969) has been a leading proponent in this country; a two-parameter extension of the Rasch model by Urry (1970); a slightly different two-parameter logistic ogive model developed by Birnbaum (1968); a similar model based on the normal ogive, developed by Lord (1952; Lord & Novick, 1968); and a three-parameter logistic ogive model (Birnbaum, 1968). All of these models express the probability of a correct (or keyed) response to a dichotomously scored test item as an ogive function of attribute level. Syntactically, this may be expressed

$$P\ (1/A) = F\ (a,b,c;A). \qquad (1)$$

The expression on the left of the equality is the probability of the keyed (1) response to item g, given A, the attribute level. F (a,b,c;A) is a general mathematical function in the item parameters $\underline{a}$, $\underline{b}$, and $\underline{c}$ and the person parameter, attribute level $\underline{A}$. In the ogive models, F is an ogive function of the distance ($\underline{b}$ -$\underline{A}$), a scale parameter $\underline{a}$, and an asymptote parameter, $\underline{c}$.

Where more than one item is administered, the probability of any pattern ($\underline{V}$), or vector, of item scores may be calculated readily by virtue of a local independence assumption. Thus

$$P\ (v/A) = \prod_{g\alpha 1}^{k} [P\ (1/A)]^{u_g} [1-P\ (1/A)]^{1-u_g}. \qquad (2)$$

Here P (v/A) is the probability of the pattern of item scores (1's and 0's), given A; u_g is the dichotomous score on item $\underline{g}$. From P ($\underline{v}$/A) we may derive expressions for the likelihood of any given attribute level, given the item response vector. This permits us to apply statistical techniques to the estimation of A, if the response pattern, $\underline{v}$, and the item parameters are known (or estimated) beforehand. There are also simple, nonstatistical techniques for combining item responses into other indices of individual differences on the attribute. (See Lord, 1974a, for pertinent discussion.)

Given that latent trait test theories in principle can satisfy the special requirements of adaptive tests, it remains to explicate such theories sufficiently to provide practical methods for estimating the parameters of each test item's characteristics curve and for estimating examinee location on the attribute scale.

State of the Art

Statistical methods for estimating item parameters and attribute levels have been developed for all the ogive models mentioned above. Computer programs for item parameter estimation are available (commercially or by private arrangement) from sources listed in Table 1. Most of these computer programs perform simultaneous estimation of examinee "ability" and of the item parameters. The statistical estimation techniques used by these programs range from simple approximations in FORTAP (Baker & Martin, 1969) to maximum likelihood in LOGIST (Wood, Wingersky, & Lord, 1976), FORTAP and BICAL (Wright & Mead, 1977), to Bayesian model estimation in OGIVEIA (Urry, 1976).

Table 1

Existing Computer Programs for Estimating Item Parameters of Latent Trait Item Response Models

Response model	Program name	Available from
1--parameter logistic (Rasch model)	BICAL	B. Wright, U. of Chicago
2--parameter logistic	LOGOG	R. D. Bock, U. of Chicago
2--parameter normal ogive	FORTAP NORMOG	F. B. Baker, U. of Wisconsin R. D. Bock, U. of Chicago
3--parameter logistic	LOGIST	R. M. Lord Educational Testing Service
3--parameter logistic	OGIVEIA or ANCILLES	V. W. Urry Office of Personnel Management

Item parameter estimation procedures generally entail simultaneous estimation of a person's ability. The task of ability estimation (or test scoring) in the context of adaptive testing is less demanding. All item parameters have been estimated beforehand; what remains is to estimate ability (or to score the tests in some other appropriate way) from knowledge of the item responses and the item parameters. The state of the art of scoring adaptive tests is outlined below.

To summarize, latent trait theories have been shown to provide appropriate psychometric bases for adaptive testing (see Lord, 1974a; Urry, 1977). These theories have been well explicated for application

to tests of unidimensional attributes, using dichotomously scored items. Mathematical algorithms have been developed for scaling attribute variables and for estimating item characteristic curve parameters and examinee ability or attribute level. These algorithms have been incorporated into computer programs that process raw item responses and yield the desired parameter estimates. These computer programs are available from their developers.

Generalizations of latent trait methods to measure unidimensional variables by means of nondichotomous test items have also been accomplished. Samejima (1969) presented methods for extending the normal ogive response model to graded response items. She has since extended it to apply to items having continuous responses (Samejima, 1973). Bock (1972) developed equations for estimating item parameters and individual ability from nominal category responses to polychotomous test items. Although they have seen relatively few applications, Samejima's and Bock's algorithms have been incorporated into available computer programs. Using graded, polychotomous, or multinomial-response test items has potential for appreciable gains in psychometric information[1] compared to the information in dichotomously scored items.

A further advance in latent trait item response models is the extension of these models to handle multidimensional test items. Samejima (1973) has begun work in this area, as has Sympson (1977).

The Design of Adaptive Tests

Discussion

Choosing an Adaptive Testing Strategy. An adaptive test is one that tailors the test constitution to examinee ability or attribute level; given this definition, we are confronted with the problem of how to accomplish tailoring. This problem of individualized test design can be brought into conceptual focus by considering that, given a fixed large set of test items from which only a relatively small subset is to be administered to an individual examinee, there exists a subset that is optimal, in some sense, at any specified test length. The items that constitute the optimal subset will vary as a function of the individual's attribute level. The problem of adaptive test design is that of selecting approximately optimal item subsets for each individual examinee. Solutions to this problem are called strategies for adaptive test design.

An adaptive testing strategy consists, minimally, of rules for item selection and for test termination; a scoring procedure may also be an integral part of some strategies. For comprehensive reviews of

---

[1] The term "information" here refers to information in the sense presented by Birnbaum (1968) and discussed below.

a variety of adaptive test strategies, see Weiss (1974) or Weiss and Betz (1973).

The essential rationale for adaptive item selection involves administering more difficult items following successful performance and easier items following less successful performance. If the test is item-sequential, this translates to selecting a harder item after a correct item response, and an easier item following an incorrect response. Choosing the appropriate difficulty increment is one aspect of the design problem. Another central aspect is choosing the criterion to be optimized.

The purpose of mental testing usually is to order examinees with respect to their relative attribute status. To achieve this purpose, it is necessary to be able to discriminate accurately between any two examinees, no matter how close they are in terms of the attribute. The required discriminability has implications for the traditional difficulty index of the items to be chosen: Using dichotomous items on which guessing is no factor to discriminate best about a point, choose test items for which the probability correct is .50 at the point in question. If guessing is a factor, the optimal p-value will exceed .5 by an amount that is a function of the effect of guessing. However, if the available test items also differ with respect to discriminating power, the latter also must enter into the determination of which item discriminates best locally. The information function (Birnbaum, 1968) of a test item provides a single numerical index by which test items may be ordered with respect to their usefulness for discriminating at a given point. In terms of equation, the information I in item g at attribute level A is expressed as

$$I_g(A) = \left[\frac{\partial / \partial A \; P_g(1/A)}{\sqrt{[P_g(1/A)][1-P_g(1/A)]}} \right]^2 \qquad (3)$$

That item is "best" for which the local value of $I_g(A)$ is highest. For a k-item test, the best subset of k items is the subset for which $I_g(A)$ is locally highest. The implication for adaptive test design is to choose items so as to maximize $I_g(A)$ at all points A. This maximization is the goal of adaptive test design. Adaptive testing strategies may or may not explicitly seek to achieve this goal; and the goal may be realized to a greater or lesser extent by the different test strategies.[1]

---

[1] Analogous to the item information function are two others--the test information function and the test score information function, both of which index measurement precision as a function of attribute level.

Adaptive test strategies differ in a number of ways. One general dimension of these differences is their item selection mode. Some strategies arrange test items a priori by difficulty and discrimination into a logical structure, such as a one- or two-dimensional matrix, and select items sequentially according to examinee performance by branching to a predetermined location in the structure and administering the item(s) that reside in that location. Such strategies may be called "mechanical" by virtue of their almost mechanical rules for item selection. Examples of mechanical strategies include the simple branching strategies; the stair-step or pyramidal method used by Bayroff and Seeley (1967) and by Larkin and Weiss (1974) and described by Lord (1974a); the flexilevel tailored test devised by Lord (1971a); the simple two-stage strategy, investigated by Lord (1971b) and by Betz and Weiss (1974); the stratified adaptive (STRADAPTIVE) procedure proposed by Weiss (1973); and even the Robbins-Munro procedures described by Lord (1971c; 1974a).

Distinguished from the mechanical, or branching, strategies are adaptive strategies that use mathematical criteria for item selection. Such strategies typically estimate the examinee's latent attribute status after each item response, then choose the available item from which some mathematical function of that estimate and of the item parameters is maximized or minimized. Examples of mathematical strategies include Owen's (1969, 1975) Bayesian sequential procedure, in which a quadratic loss function is minimized; and Lord's (1977) maximum likelihood strategy in which the available item with the largest local information function is chosen.

One of the clearest distinctions between mechanical and mathematical strategies is that in the latter every unadministered test item is potentially eligible for selection at any stage in the test, whereas in a mechanical strategy only a small number of items--as few as two--are eligible for selection at any given stage. Another obvious distinction is that the mathematical strategies are appealing by virtue of their elegance, whereas the virtue of the mechanical strategies is their simplicity. In confronting the problem of choosing an adaptive strategy, one first must choose between elegance and simplicity. Then, by electing categorically either a mechanical or mathematical strategy, one is faced with the further choice of a specific adaptive testing strategy. The number of strategies proposed for use has proliferated faster than have research results useful to guide the choice.

The Test Length Issue. Confounded with the problem of choosing a testing strategy is the problem of test length. Like conventional tests, adaptive tests may be short or long; unlike most conventional tests, adaptive tests may adapt test length, as well as test design, to the individual.

The notion of variable length test seems to make sense, since the examiner can administer as few or as many items as necessary to measure each individual with a specified degree of precision. Furthermore, it

is apparent that if measurement precision is to be held constant, achieveing that precision should require relatively few items for persons whose attribute level is near the central tendency of the population, and more items for persons located in the upper and lower extremes of the attribute continuum. Roughly speaking, if precision is to be held constant, the required adaptive test length should be a U-shaped function of attribute level.

Among the proponents of variable length adaptive tests are Samejima (1977), Urry (1974, 1977a), and Weiss (1973). Weiss advocates the use of a simple stopping rule based on identifying a "ceiling level" of difficulty for each examinee in conjunction with stratified adaptive (STRADAPTIVE) strategy. Samejima (1977) proposed that test length be varied such that a constant level of measurement precision (indexed by the test information function) be achieved throughout a prespecified range on the attribute scale. Urry (1974) espouses using variable test length in conjunction with Owen's Bayesian sequential adaptive strategy in such a way as to yield a prespecified level of the validity[1] of the test scores as a measure of the underlying attribute; the squared validity may be interpreted as a reliability coefficient.

It should be pointed out that some adaptive testing strategies are inherently fixed-length. Among these are the flexilevel, pyramidal, and two-stage strategies. Others, like Weiss' and Owen's strategies, make fixed-length optional. The variable-length test termination criteria espoused by Urry and Samejima can in principle be used with any adaptive strategy--even the ones described above as inherently fixed-length. Weiss' criterion for variable-length termination of the STRADAPTIVE test, however, is somewhat restricted in applicability because it requires a certain structure--stratification by difficulty--of the item pool.

Given the intuitive appeal of variable test length, two problems remain. One problem is to decide between variable versus fixed test length and which of the available test termination criteria to adopt. The other problem is to verify that the apparent advantages of variable test length are realized in practice.

State of the Art

Choosing an Adaptive Strategy. One of the first steps in implementing a program of adaptive testing must be to choose an adaptive testing strategy from among those available. This choice should be an informed one, based on the results of research comparing the merits

---

[1] By "validity" is meant the correlation between the test score (ability estimate) and the underlying true ability. This correlation is estimated from the Bayesian posterior variance under Owen's method following each item response by an examinee.

of available methods. Very little research has been conducted along these lines, however. Instead, most adaptive testing research has concentrated on comparing the psychometric properties of specific adaptive test strategies against the properties of otherwise comparable conventional test designs. Weiss and Betz (1973) reviewed the results of these comparisons.

Some live-testing research comparing adaptive strategies was reported by Larkin and Weiss (1975). Only two strategies were compared, however, and the results were equivocal. The only other data available as a basis for comparing adaptive strategies are data resulting from analytic studies of the properties of various strategies and from model-sampling computer simulation studies of similar properties. Lord (1970; 1971a, b, c) reported the results of analytic studies of several adaptive strategies, but made no effort to compare them. The only studies that directly compared several strategies were the simulation studies of Vale (1975) and McBride (1976b).

Vale's study compared five leading strategies in terms of the level and shape of the resulting test information functions; in other words, in terms of relative measurement precision as a function of attribute level. Vale's artificial data were based on a response model that did not permit guessing. Further, he presented data only for 24-item fixed-length tests. His results indicated that under the conditions simulated, the Bayesian test strategy was superior in terms of the level of measurement precision, whereas the stradaptive strategy was superior in terms of measuring with constant precision at all levels of the attribute. The other adaptive strategies compared--the flexilevel, pyramidal, and two-stage strategies--all were inferior to the first two in some way.

Vale's study simulated only the no-guessing situation and a single test length and did not investigate mathematical strategies other than the Bayesian one. McBride (1976b) extended Vale's results in a series of simulation studies comparing the psychometric properties of two mathematical and two leading mechanical strategies at six different test lengths and under several realistic conditions, including the presence of guessing. His results indicated that the two mathematical strategies were generally superior to the mechanical ones, especially at short test lengths (5 to 15 items), both in terms of test fidelity (validity) and measurement precision. At moderate test lengths (20 to 30 items), the mathematical strategies were still superior, but their advantages over the mechanical strategies were slight.

The two mathematical strategies were Owen's Bayesian sequential one, and a variant of a maximum likelihood strategy proposed by Lord (1977). Differences in results between the two were slight, but the maximum likelihood strategy was judged superior in adaptive efficiency--the degree to which the methods select the optimal subset of items at a given test length--and also in several other respects.

McBride concluded that his data favored the maximum likelihood strategy overall, but that the choice among the four strategies should be influenced by other considerations. For example, the Bayesian strategy was the best of the four, in terms of adaptive efficiency, at very short test length (5 items) when all examinees began the test at the same level of difficulty; at the longer test lengths (25 and 30 items), all four strategies had excellent measurement properties, and any one of them could reasonably be chosen.

It is important to note that McBride's comparison studies were carried out so that the correct test item parameters were known and available when simulating each adaptive strategy. In live testing, of course, only fallible estimates of the parameters of the item characteristic curves are available. The use of fallible estimates should introduce measurement errors over and above those entering into McBride's data. It is possible that the effects of such errors could alter some of the conclusions McBride reached concerning the order of merit of the four strategies he evaluated. Research is needed extending his findings to the case of fallibly estimated item parameters.

Vale's (1975) and McBride's (1976b) simulation studies are the only ones available for comparing strategies. There is, however, a sizable body of research results available for evaluating several individual adaptive strategies against conventional tests. Urry and his associates (Urry, 1971, 1974, 1977b; Jensema, 1972, 1974, 1977; Schmidt & Gugel, 1975) have reported results of a comprehensive program of computer simulation investigations of some psychometric properties of Owen's Bayesian sequential adaptive test. Vale and Weiss (1975) report in considerable detail the measurement properties of the stradaptive strategy. Lord (1977) recently proposed the broad-range tailored test (a maximum likelihood strategy) and reported some data relevant to its psychometric properties. All of these investigations have utilized model-sampling computer simulation methods to explore the behavior of the various test strategies. All have also taken different lines of approach and concentrated on different aspects of each strategy's psychometric behavior, so that it is not possible to compare the strategies on the basis of the available reported data.

Fixed-Length Versus Variable-Length Adaptive Tests. There has been no systematic study of the relative merits of variable-length versus fixed-length adaptive tests. Rather, researchers in this area have tended to make an a priori choice between the two options and leave the choice unquestioned. Working independently and motivated by different considerations, Samejima (1976), Urry (1974), and Weiss (1973) all chose in favor of variable length. Lord (1977), however, opted for fixed length in proposing his broad-range tailored test.

Samejima (1977), working in the framework of a maximum likelihood strategy, suggested that the test information function be estimated for each individual after each item response. The test may be terminated when the estimated value of the information function reaches a prespecified level. The effect of using the test termination rule

would be to achieve a virtually horizontal test information function throughout a wide interval of the attribute continuum. This is tantamount to using the test termination rule to guarantee equiprecision of measurement over a specified range, which is one of the principal motivations behind adaptive testing. No data are available to indicate whether Samejima's test termination criterion would actually achieve its purpose.

Urry and others (Urry, 1974; Jensema, 1977; Schmidt & Gugel, 1975) favor variable test length for use with Owen's Bayesian sequential adaptive test strategy. Under Owen's procedure, the posterior variance of the distribution of the Bayes estimator is calculated following each test item response; that variance, which usually diminishes after each item, is interpreted by Urry as the square of the standard error of estimate (s.e.m.) of the examinee's attribute level. Thus, by terminating each test when the calculated standard error reaches a prespecified small value, the standard error of estimation in the examinee group can be controlled and consequently so can an index of reliability of the ability estimates. Thus, Urry advocates a variable length test termination rule to ensure (approximately) that the adaptive test scores have a prespecified level of correlation with the latent attribute being measured.

Urry (1971, 1974) and Jensema (1977) have presented the results of numerous simulation studies of Owen's procedure to show that the fidelity coefficient of the test scores can be controlled by using the posterior variance as a test termination rule. These studies all used the true values of the simulated test items' parameters for item selection and scoring. Schmidt and Gugel (1975) presented simulation study data for the more veridical case in which fallible item parameter estimates are used. The effect of using fallible item parameters with Owen's procedure was a tendency for the tests to terminate prematurely, with the result that the obtained fidelity coefficients fell slightly short of the targeted values.

Subsequent to his computer simulation studies, Urry (1977b) administered Bayesian adaptive tests of verbal ability to live examinees. His analysis of the adaptive test data evaluated the usefulness of the s.e.m. test termination criterion for controlling the level of "construct validity"--correlation of the resulting test scores with an independent measure of the same ability. Urry found that for all the evaluated levels of the s.e.m. criterion, the obtained validity coefficient was equal to or slightly greater than the forecast validity associated with each test termination criterion. He concluded that the theory was supported; that it was possible to control the reliability and validity of a test by using the Bayesian procedure and manipulating the posterior variance termination criterion.

Urry and others have been successful in controlling adaptive test validity/fidelity/reliability by manipulating test termination criteria. That success notwithstanding, they have not demonstrated that equiprecision of measurement (a flat information function) could be achieved

using their proposed variable test length procedure, nor have they attempted to do so. McBride (1977), in simulation studies of the same Bayesian procedure, found a strong positive correlation (.8) between test length and ability when variable test-length was used; i.e., the termination criterion was satisfied in fewer items for lower ability examinees. His data indicated that the relationship between attribute level and test length is not U-shaped, as it should be to approximate a horizontal information function. As a result, the information functions of the simulated Bayesian variable-length tests tended to be convex in shape, with markedly low values in the low end of the attribute range. McBride concluded that there may be greater virtue in fixed-length Bayesian adaptive tests.

It should be clear by now that some issues involved in choosing an adaptive testing strategy and in deciding between fixed and variable test length remain unresolved. Additional research in both areas is needed.

These unresolved issues need not impede progress in the experimental implementation of systems for adaptive testing, because the unknown differences among the leading adaptive testing strategies are undoubtedly of lesser magnitude than the difference between any such strategy and a conventional test design. Perhaps recognizing this, Urry (1977a) cautions sternly against procrastination in implementing adaptive testing. It may be wiser to proceed by making a tentative choice among the strategies and an arbitrary decision on the test length issue, letting the academic world settle the remaining basic research issues in due course.

Scoring Adaptive Tests

Discussion

For most adaptive test strategies, the traditional number correct or proportion correct score will not suffice to index individual differences on the attribute being measured. To understand this, consider the goal of adaptive testing: to achieve equiprecision of measurement across a wide range. The goal is achieved by fitting the test to the examinee. Other things being equal, accomplishing that fit will result in a flat regression of the proportion correct score on the attribute scale. That is, test difficulty (as indexed by mean proportion correct) will be approximately equal across a wide range of the attribute. As a result, the proportion correct scores will have an information function whose value is near zero throughout that wide range (e.g., McBride, 1975).

In practice, adaptive tests can be expected to fall somewhat short of the goal of equiprecision, so that there may be some information in traditional scoring methods. Nonetheless, for the most part the proportion correct and similar indices are not adequate as general scoring procedures for adaptive tests.

An immediate exception is Lord's (1971a) flexilevel test strategy, which was specifically designed so that the number correct score would be a meaningful index. The flexilevel strategy aside, let us consider the requirements and desiderata of an adaptive test scoring procedure. In an adaptive test, different persons take different sets of test items. These items vary in difficulty and may also vary in their discriminating powers and susceptibility to guessing. Further, under some adaptive strategies, test length may vary from one person to another, as may the difficulty level at which the test was begun. There is useful information in all the parameters just mentioned, so that a scoring method needs to account not only for how many items a person answers correctly, but also which items were answered, and in some cases which ones were answered correctly or incorrectly. It is desirable for the scoring procedure to make use of all the information contained in the examinee's answers, as well as in the identity of the items constituting the test.

Scoring methods based on latent trait theory are especially useful and appropriate for scoring adaptive tests. This is because such methods can take into account all relevant data in the constitution of an individual test--such as test length and item characteristic curve parameters--as well as the item-by-item performance of the examinee. Some fairly simple methods are available, along with others so complex that they require a computer to perform needed calculations. The problem of scoring adaptive tests is the problem of choosing (or devising) an appropriate scoring method. Some of the available methods are discussed below.

State of the Art

The number of scoring methods available for adaptive tests is sizable. Some methods are general and are applicable under a variety of testing strategies, while others were devised ad hoc and are specific to one or a few strategies. Among the general methods we can distinguish statistical procedures from nonstatistical ones.

Statistical Scoring Procedures. These procedures are based on techniques of combining known psychometric information about the test items with the observed item response performance of the examinee in such a way as to yield a statistical estimate of the examinee's location on the attribute scale. Although there are a host of such estimation methods available, the ones most prominent in the literature have been estimators based on the Rasch one-parameter logistic ogive item response model, on the Birnbaum three-parameter logistic ogive model, and on the three-parameter normal ogive model.

Under the Rasch model, the number correct score is a sufficient statistic for the estimation procedure,[1] provided that the Rasch difficulty parameters of the items constituting an individual test are known, there is no guessing, and all items are equidiscriminating. Least squares estimators and maximum likelihood estimators of attribute level have been derived and published (e.g., Wright & Panchapakesan, 1969). The maximum likelihood estimator is somewhat more elegant and more accurate. Estimators based on the Rasch model are not strictly appropriate for scoring tests having known differences in item discrimination parameters, or on which there is a substantial chance of answering questions correctly by guessing. Urry (1970) has evaluated the effects of ignoring guessing and item discriminating powers in scoring adaptive tests; the result is some loss of accuracy in ordering individual differences. That loss is reflected in the validity of the adaptive test scores for measuring the relevant attribute. In sum, where the Rasch model is appropriate, its use for scoring adaptive tests is not questioned. Where it is inappropriate, a scoring procedure based on a more general response model will extract more useful information from adaptive test response protocols.

The more general item response models include two- and three-parameter normal and logistic ogive models. The logistic models can readily be made to approximate closely the normal models. Because of their mathematical tractability, the logistic ogive models have largely supplanted the normal ogive models in use. Further, the three-parameter models are more general, of which the two-parameter ones are special cases; similarly, the Rasch model is a special case of three-parameter logistic model. Thus, the three-parameter logistic model is the model predominantly used in current practice.

Test scoring (attribute estimation) under the three-parameter logistic model usually has been accomplished using iterative maximum likelihood estimation procedures. Such procedures use all the information available in an examinee's dichotomous item scores on an adaptive (or conventional) test: item difficulty, discrimination, and guessing parameters; and the pattern of the examinee's right and wrong answers. The likelihood equations used for this scoring method have been derived and published (e.g., Jensema, 1972). Algorithms for performing the estimation procedure have been incorporated in several computer programs (e.g., see Urry, 1970; McBride, 1976b; Wood, Wingersky, & Lord, 1976; Bejar & Weiss, 1979).

Methods other than maximum likelihood may also be used for the statistical estimation of attribute scale location. Sympson (1976), for example, recently described two alternative methods, including a

---

[1] For scoring an adaptive test using the Rasch model, the number correct is not admissible as a test score, but rather as a sufficient statistic for estimating ability; the resulting estimate is the test score.

generalized Bayesian one, for estimation under the three-parameter logistic model.

There is one prominent application of the three-parameter normal ogive response model to estimating examinee location on the attribute scale, a Bayesian sequential procedure given by Owen (1969, 1975). Owen's estimation technique was presented as an integral part of his sequential adaptive testing strategy. It is just as appropriate for use as a scoring procedure for any test where item parameters and dichotomous item scores are available.

Both the maximum likelihood procedure and Owen's Bayesian sequential procedure are methods of estimating an examinee's location on a continuum. There are substantial differences in approach between the two, however. The maximum likelihood procedure estimates the examinee location parameter from the pattern of an examinee's right and wrong answers to his or her test questions, by solving a likelihood equation. No prior assumptions are involved regarding the examinee's location or the distribution of the attribute.

Owen's Bayesian procedure estimates examinee location sequentially. It begins with an initial estimate of the location parameter and updates that estimate, one item at a time, by solving equations that consider both the likelihood function of the single item score and the density function of an assumed normal distribution. The ability estimate is the final updated value after the last item score is considered.

Because it is a sequential procedure, Owen's scoring method is order-dependent. Analyzing the same item responses in different orders can result in slightly different numerical values of the final estimates. The maximum likelihood scoring procedure is not dependent on the order in which items are administered (or item responses analyzed).

Another noteworthy difference between these two methods concerns their statistical properties. Owen's Bayesian estimator behaves like a regression estimate: Extreme values are biased toward the initial (prior) estimate, which is the mean of the normal Bayesian prior distribution assumed for the location parameter. This bias may not be linear, as McBride (1975) demonstrated, and may be undesirable for applications (such as criterion-referenced testing) in which the numerical accuracy of the estimator is of some consequence. Urry (1977a) pointed out that the bias in the Bayesian estimates is readily correctable using an ancillary method, but no data are available concerning the efficacy of Urry's proposed correction. The maximum likelihood estimator does not seem to be subject to the systematic bias of Owen's Bayesian scoring method, but requires appreciably more computer processing time and sometimes fails to converge to a satisfactory estimate (McBride, 1975).

Sympson (1976) reported developing two alternative methods for the examinee parameter estimation problem. One method is a Bayesian method that considers the examinee's entire vector of item scores at

once and thus avoids the order-dependence of Owen's sequential scoring method. It is also more general than Owen's method in that it is not restricted to assuming a normal prior distribution on the latent attribute. Instead, the user is free to specify any form for the Bayesian prior distribution.

Nonstatistical Scoring Procedures. The scoring methods discussed yield statistical estimates of an examinee's location on a scale. Several less-sophisticated scoring methods are available that yield numerical indices useful for ordering examinees. Such methods have the advantage of computational simplicity, but lack the properties of statistical estimators. Indices have been proposed for several different adaptive testing strategies. Some of these indices are specific to the strategies that gave rise to them, while others are generalizable to two or more adaptive strategies. Weiss and Betz (1973) and Weiss (1974) have discussed nonstatistical scoring methods in detail. Vale and Weiss (1975) evaluated alternative methods against one another and found one originally proposed by Lord to be generally superior to the others. This index, called the "average difficulty score," is computed by summing the item difficulty values of all test items answered by an examinee and computing the average. The item difficulty values involved are the difficulty parameters of the item characteristic curves, not the traditional p-value difficulty indices.

The average difficulty score is appropriate for adaptive tests in which all examinees begin testing at the same difficulty level. Although it may be used in conjunction with tests having variable entry levels, its properties have not been systematically investigated in such a context. The weight given to the difficulty of the first item in a variable entry level test may have the effect of biasing test scores in the direction of the pretest estimate of the examinee's ability.

An alternative to the average difficulty score is to calculate only the average difficulty of the items answered correctly; however, test scores calculated in this fashion correlate almost perfectly with the average difficulty of the items administered (Vale & Weiss, 1975). Other nonstatistical scoring procedures evaluated to date have been generally inferior to these two methods, even for scoring appropriate types of adaptive tests; therefore, they will not be discussed here.

The Testing Medium

The adaptive test merits consideration as a possible replacement for conventional standardized group tests. Therefore, the test administration medium must be amenable to testing relatively large numbers of examinees. There is a need to identify media that can meet this requirement and to evaluate such media both absolutely and in a comparative sense.

The media available for administering adaptive tests fall into two categories: specially designed paper-and-pencil tests and automated testing terminals. A paper-and-pencil adaptive test superficially resembles a conventional test, but requires the examinee to comprehend and follow relatively complex instructions for the sequential choice of test items and for marking item responses. The added complexity of the examinee's task in taking a paper-and-pencil adaptive test may be excessive, particularly for lower ability persons, with the result that the dimension to be measured is confounded with the examinee's ability to follow the instructions. If such a confounding occurs to any substantial degree, the test may be an invalid measure of the intended trait dimension. An obvious research issue is to inventory the available methods for administering adaptive tests in the paper-and-pencil medium and to evaluate the extent to which examinee task complexity is excessive.

Automated administration of an adaptive test relieves the examinee of the burden of complying with the complex instructions; instead, the testing device assumes this burden. This benefit is not achieved without cost, however. Typically, automated tests have been administered at interactive computer terminals, a medium currently more expensive than paper-and-pencil administration. For adaptive administration of tests composed of items like those in paper-and-pencil group tests--typically, multiple-choice items--in principle, a device much less sophisticated than a CRT computer terminal will suffice. Test administration using such a device should be considerably less expensive than the use of a computer. Clearly, the identification and design of alternative devices for automated testing is an important issue for research and development.

State of the Art

Paper-and-Pencil Adaptive Tests. Bayroff, Thomas, and Anderson (1960) designed experimental paper-and-pencil branching tests based on Krathwohl and Huyser's (1956) scheme for a "sequential item test," a pyramidal adaptive strategy (Weiss, 1974). On subsequent administration of branching tests of word knowledge and arithmetic reasoning, respectively, Seeley, Morton, and Anderson (1962) found that 5% and 22% of the examinees made critical errors in following the item branching instructions. Such errors made those examinees' answer sheets unscorable under the scoring method used; the tendency to such errors was related to general ability.

Lord (1971a) devised the flexilevel testing method, an adaptive strategy specifically intended for paper-and-pencil testing. Olivier (1974) administered flexilevel tests of word knowledge to 635 high school students and found that 17% of his examinees' tests were unscorable because they had made critical errors in branching.

The Seeley et al. (1962) and Olivier (1974) experiences have created an air of pessimism about the feasibility of using the paper-and-pencil medium for adaptive testing. This pessimism is based on two facts: (a) A substantial proportion of examinees tested has been unable to follow the item-to-item branching instructions; (b) under the scoring methods used, certain branching errors made the tests unscorable. If the concept of paper-and-pencil adaptive tests is to be salvaged, both problems must be solved. That is, the complexity of the examinee's task must be reduced, and scoring procedures must be devised that can accommodate item branching errors.

The statistical scoring methods based on item characteristic curve theory, discussed in another section, satisfy the latter requirement. They provide a means of calculating a score, using a common metric, for examinees who answered different sets of test items. These scoring methods are applicable even to examinees who erred in item branching, provided that it is known which items were answered and whether the answers were right or wrong.

Since the use of item characteristic curve theory in effect solves the scorability problem, all that remains to make paper-and-pencil adaptive testing feasible is to minimize the problem of the complexity of the branching task. This problem has not been solved to date, although tentative approaches to its solution have been taken (e.g., McBride, 1978).

Perhaps the simplest solution proposed is the "self-tailored test" suggested by Wright and Douglas (1975) for use with test items that satisfy the Rasch simple logistic response model. Test items are printed in the booklet in ascending order of difficulty. The examinee is instructed to start answering test items at whatever difficulty level he or she chooses and to stop where he or she chooses (or perhaps to answer a fixed number of items). The test score (a Rasch ability estimate, which can be determined by referring to a preprinted table) would be a function of the difficulty levels of the easiest item answered and the most difficult item answered, and the number of items answered correctly in between.

The Wright and Douglas notion is appealing in its simplicity, but it has drawbacks. First, its psychometric merits depend heavily on the ability and willingness of the examinee to choose test items that are most informative for ability level--neither too difficult nor too easy. Second, its linear branching rules and ability-estimation procedures are not strictly appropriate where guessing is a factor and where there is appreciable variability in the discriminating powers of the test items. Nonetheless, this "self-tailored" testing scheme is worthy of some exploratory research in settings where it is desirable to reduce substantially the number of items each examinee must respond to.

Where guessing is a factor and items vary appreciably in discriminating power, the optimal choice of items in an adaptive test is a function of those variables as well as of item difficulty. This suggests

that a somewhat more sophisticated rationale is required for adaptive item branching than the simple linear progression implicit in the Wright and Douglas proposal. Implementing a true item branching procedure in a feasible paper-and-pencil version, without overbearing complexity, may call for new approaches. The necessary approach is to minimize the opportunity for error by making the branching instructions as simple as possible and as few as possible.

Simplicity may be achieved by using latent ink technology in designing and printing answer sheets, thereby making the branching instruction unambiguous and contingent only on what answer the examinee gives to the item he or she is currently working on. The frequency of item branching can be reduced by using a modified two-stage adaptive strategy; the first stage might be a short branching test of several items, while the second stage might be a multilevel test. The function of the first stage test would be to route the examinee to an appropriate level in the second stage. Each level would have the format of a short conventional test; thus, no branching instructions need be followed during the second stage. This notion was developed further in a separate paper (McBride, 1978).

Automated Adaptive Testing. Most research on adaptive testing has focused on computers as control devices and on computer terminals as the medium for test administration. Although the computer is a convenient and apt tool for automating testing, the relationship of computers to adaptive tests is sufficient but not necessary. Any device capable of storing and displaying test items, recording and scoring responses, and branching sequentially from item to item can in principle suffice as the testing medium. The computational power of a computer may be highly desirable for implementing some adaptive testing strategies, but it is far from necessary for all. Further, tests based on dichotomously scored multiple-choice test items make such minuscule demands on the capability of a modern computer that use of a computer solely for administration of such tests seems wasteful. Simpler and less costly devices can do the job, and such devices should be developed.

The first concrete effort to develop a simple device for automated adaptive testing seems to have been one made at the Air Force Human Resources Laboratory, Technical Training Division (AFHRC/TT). Personnel there have developed a prototype programmable microprocessor terminal for administering an adaptive test (Waters, personal communication). The terminal itself resembles a hand-held desk calculator, with an array of numbered keys used to respond to test items. Its display device is a small array of several light-emitting diodes (LEDs). The unit is preprogrammed to direct an examinee to answer a response-contingent sequence of test questions that are printed in a separate test booklet. After recording and scoring the examinee's response to the current test item, the microprocessor unit computes the location of the next item; the LED displays that location as an item number; the examinee then turns to that item in the test booklet and responds by keying in an answer on the keyboard. At test termination, the examinee's protocol of identification data, item responses, and test score can be

"dumped" to a special-purpose computer before the next examinee is tested. Development of the AFHRL prototype is being undertaken by an independent contractor.

A direct extension of the microprocessor concept is contemplated by AFHRL/TT. This would involve using the programmable microprocessor both for item selection and for controlling the display on a peripheral device of test items stored in microform: film slides, microfilm, or microfiche. The contemplated device would emulate the function of a full-scale computer terminal, but with limited interactive capability. The significance of this step is that the examinee's role would be limited to answering the sequence of displayed test items; the examinee would not have to participate in item selection or in locating selected items.

In considering the state of the art with respect to automated testing terminals, it is useful conceptually to consider the separate components required of a test delivery device. These include the following:

- Stimulus/display device
- Response device
- Item storage medium
- Internal processing
- Response processing capability
- Item selection capability
- Test scoring capability
- Data recording capability.

Display devices proposed or in use range in complexity from simple printed matter, to microform readers, to computer graphics terminals. Microform readers include microfilm reel readers, manual microfiche readers, and automated magazine microfiche and ultrafiche readers. These microform devices are capable of storing and displaying any test material that can be printed and photographed, including graphic material. The computer terminals amenable to automated testing include teletypes, monochrome CRT terminals, plasma tube (PLATO) terminals, and color graphics CRT terminals. Computer terminals typically have integral provisions for response keyboards; microform display units do not. All devices listed above are commercially available off the shelf; special provisions may be required to integrate each into a testing system and to interface each to a test control device.

With CRT or similar computer terminals, test item storage must be in computer code, either core-resident or mass storage resident and rapidly accessible. The volume of displayable material needed to support a full battery of adaptive tests may require hundreds of thousands of characters of computer storage.

Microform storage of test items is more efficient but less flexible than computer storage. Items may be photographed and stored on

microfilm rolls, photographic slide magazines, microfiche, or ultrafiche. Slide magazines are bulky and cumbersome and worth considering only as a prototype. Microfilm rolls are a highly efficient storage medium, but the machinery needed to implement adaptive testing with items stored on microfilm is expensive and inappropriate. Microfiche and ultrafiche seem to offer an acceptable compromise. A single 4-by-6-inch microfiche can contain several hundred display images; an ultrafiche of similar dimensions can hold about 2,000 images. Thus, test items for a sizable battery of adaptive tests could be stored on about ten microfiche or on a single ultrafiche. All that is required for a test item display device is the ability to automate the microfiche/ultrafiche reader.

Automated microfiche readers are already commercially available and can be modified readily to serve as testing terminals by interfacing them to appropriate control devices.

The internal processing requirements of automated adaptive testing may be accomplished by a central computer, minicomputer, or microcomputer, entirely within today's state of the art. System design stands between current development and implementation of a computerized system for adaptive testing.

Some efficiency or cost effectiveness may be gained by the use of special-purpose microprocessors to control the test itself and the testing equipment. Again, such devices are well within the current state of the art in electronics. The equipment needs to be designed and integrated into a system for adaptive testing.

Item Pools Development

Discussion

Adaptive testing involves selective administration of a small subset of a larger pool of items that measure the trait of interest. The size of this item pool, along with the psychometric characteristics of the constituent items, places limits on the measurement properties of the adaptive test. Obviously, the item pool should be large enough and constituted so as to permit the adaptive tests to function effectively. Early theoretical research in adaptive testing suggested that item pools had to be large, ranging from one or two to several hundred or several thousand test items. More recently, computer simulation research by Jensema (1977) and other associates of Urry has shown that adaptive tests can function very well at test lengths of 5 to 30 items and that item pools containing 50 to 200 items are of sufficient size, provided that prescriptions for the psychometric characteristics of the test item are met. These prescriptions concern the magnitude of the items' item response model discrimination parameters, the range and distribution of the item difficulty parameters, and the susceptibility of the items to random guessing.

Urry (1974) has listed such prescriptions for items calibrated (with the three-parameter ogive model) against an ability scale on which the examinee population is distributed normal (0,1). They include item discrimination parameters exceeding .80, item guessing parameters below .30, and a rectangular distribution of item difficulties ranging approximately from -2 to +2 units on a standard deviation scale. McBride (1976b) suggested an even wider range of item difficulty and found that item pools with 100 and 150 items supported satisfactory measurement properties in their adaptive tests. For measurements focusing on the trait scale interval between -2 and +2 standard deviations about the population mean, a 100-item pool seems sufficient (e.g., Schmidt & Gugel, 1975; McBride, 1976b). For measurement over a wider interval, a wider span of item difficulty is indicated, along with a proportional increase in item pool size; see Lord (1977) and McBride (1976b) for examples.

Because of the requisite size of item pools for adaptive testing and the prescriptions concerning the needed psychometric characteristics of the test items, a question of the feasibility of assembling adequate item pools arises. Large numbers of test items used in conventional tests will not meet the discrimination parameter criterion for inclusion in adaptive test item pools. Furthermore, the wide, rectangular distribution of item difficulty specified by Urry's prescription may be difficult to satisfy. In many settings it may not be feasible to construct adaptive test item pools from off-the-shelf test items. However, where large-scale testing programs are already in progress, the outlook is better. Urry (1974), for example, was able to assemble a 200-item pool for adaptive testing of verbal ability by screening about 700 items in 15 forms of a U.S. Civil Service Test. Lord (1977) has made available for research a pool of 690 verbal items from obsolete forms of several tests published by the Educational Testing Service.

In military testing, current and obsolete test batteries in the aggregate contain hundreds of test items for each of several cognitive ability variables that have been measured by military tests for several years. For example, test variables such as word knowledge, arithmetic reasoning, and general information have been included in Army selection test batteries[1] through several generations of tests and multiple forms within each generation. Such tests can be expected to contain, in their various alternate forms, sufficient numbers of test items from which to select the items to constitute item pools for adaptive testing.

For test variables not having a large bank of items already in existence, a major item-writing/item-pool development program will be necessary. Even for variables already well represented in large numbers of test items, other problems remain to be solved before the

---

[1] Examples include the Armed Forces Qualifying Tests (AFQT), the Army Classification Battery (ACB), and the current Armed Forces Vocational Aptitude Battery (ASVAB).

adaptive testing item pools can be assembled. Reference here is to the problem of item calibration--estimating the latent trait response model parameters of each item's characteristic curve.

In a previous section, the existence of computer programs for estimating item parameters was mentioned. The basic data required by such programs are the dichotomously scored responses of examinees to a moderately large number of test items. Urry (1977a) and Schmidt and Gugel (1975) have reported research results that suggest that the number of examinees should equal or exceed 2,000 in order to achieve accurate estimates of the item parameters for a three-parameter item response model. Presumably somewhat smaller numbers will suffice for the simpler but less general one- and two-parameter response models. The important point is that errors of parameter estimation will increase as either or both of the two sample sizes--items and persons--decreases.

In calibrating the test items of large-scale testing programs, such as ACB and ASVAB, access to adequately large examinee samples should not be a problem, since hundreds of thousands of examinees take each form of a battery annually. However, the item sample sizes are in many cases inadequate by Urry's standards. For example, the longest subtest in the current ASVAB is only 30 items. Most ASVAB subtests are shorter. If accurate item calibration is not possible using the existing answer sheets from such subtests, then item calibration studies will need to include administration of longer subtests to large numbers of examinees in a testing program separate from current operational testing. On the other hand, if a means can be found that will permit accurate item calibration based on item responses to current subtests, there will be a substantial reduction in the expense and effort required to assemble adaptive testing item pools.

State of the Art. For estimating item parameters under a three-parameter response model, two existing computer programs are appropriate: OGIVEIA, described by Urry (1977a); and LOGIST, described by Lord (1974b). Item calibration research based on OGIVEIA led Urry to prescribe test lengths of 60 items and examinee samples of 2,000 as the minimum values for satisfactory parameter estimation. Lord (1974b) recommended a similar examinee sample size, but made no mention of the requisite test length.

Urry's program is appropriate for calibrating dichotomously scored items only; no provision is made for item scores other than right or wrong; further, it explicitly assumes a normal distribution of the ability parameter. LOGIST contains explicit provision for differentiating unanswered items from those answered incorrectly. It treats differentially two categories of unanswered items: items reached but omitted and items not reached. Items not reached are ignored during the portion of the item calibration process in which an examinee's ability parameter is estimated. Lord (1974b) has suggested that this feature of LOGIST may be useful for calibrating sets of test items in which not all examinees answer the same items. Thus it may be possible

to use LOGIST to calibrate simultaneously items from two or more alternate forms, where a different examinee sample responds to each form. LOGIST makes no assumptions regarding the form of the distribution of ability.

Two research questions need resolution before adaptive testing item pools can be constructed from existing test items. First, what are the effects of calibrating test items from the answer sheets of rather short tests (20 to 30 items)? Second, if those effects are not favorable, is it feasible to calibrate items by pooling answer sheets from two or more forms, each taken by different examinees, to increase the number of items to a size needed for satisfactory calibration? These questions are not readily amenable to answers based on theoretical or mathematical analysis. However, they may be answered empirically by means of simulated calibration of artificial item response data along lines used by Lord (1975b) or by Schmidt and Gugel (1975).

A related issue is one of equating the scales derived from independent calibrations of test items measuring a common variable but contained in different tests. This is the same problem as making item parameter estimates that result from calibration of different tests in different examinee samples all have reference to the same ability metric. Lord (1975a) has suggested a number of equating methods, based on item characteristic curve theory, that are applicable to this problem. Some of those equating methods have distinct advantages over traditional equating methods.

Advances in Measurement Methodology

Discussion

Current methods of measuring psychological traits overwhelmingly use tests composed of dichotomously scored items. In ability measurement, each such item is a task, chosen from the domain of relevant tasks, that an examinee performs successfully or unsuccessfully, correctly or incorrectly. Performance on each item task is taken as an indication of the examinee's level of functioning on an underlying ability trait. Thus, the trait is only indirectly measured, using item tasks that have only imperfect fidelity to the trait of interest. For example, multiple-choice vocabulary test items often are used to measure verbal ability.

Most adaptive testing research has used the same kinds of items. Adaptive testing using traditional item types represents an improvement in the efficiency of measurement but no improvement in the fidelity of the test behavior to the trait of interest.

The usual media of group test administration, paper-and-pencil booklets and answer sheets, necessitated the compromise of task fidelity. Administration of tests by computer terminals, as is common in adaptive testing research, opens up the possibility of introducing whole new

modes of stimulus and response to the methodology of measuring psychological abilities and perhaps of improving the fidelity between tests and abilities. The implications of computerized test administration for measurement are potentially vast, as is the number of research issues.

The basic issue is this: How can the capability of the computer be exploited to yield more and better test information about individual examinees? This subsumes other questions, such as: Can test stimuli be enriched, and/or response modes expanded, to achieve improved measures of current ability variables? Can nontraditional ability variables be identified and measured, yielding improvements in test fidelity and validity? Can advances in measurement procedures be made that are accompanied by advances in practical utility?

State of the Art

A comprehensive review of the current status of research in these issues is beyond the scope of this paper. Only a cursory overview will be attempted.

For measuring traditional ability variables, expanded stimulus and response modes are made possible by computer administration. On the response side, several different approaches are possible. One is to permit on-line polychotomous scoring rather than dichotomous scoring of traditional multiple-choice type items: Samejima (1969) and Bock (1972) have developed psychometric procedures to support such item scoring methods. A more sophisticated approach is to accept natural language, or free responses, to traditional test item stimuli; the examinees could type their answers in full on a typewriter-like keyboard rather than choose multiple-choice answers. Natural-language processing computer programs would be used to check free-form responses against the nominal correct answers and thus to score item performance (see, for example, Vale & Weiss (1977)).

Traditional test stimuli are static and usually monochrome; this is necessitated by the printed medium in use. Presenting stimuli at computer terminals makes it possible to introduce multicolored stimuli and to use dynamic test items. For example, the examinee may be permitted to "rotate" in space a three-dimensional figure presented on a CRT screen to facilitate visualization. Cory (1978) has experimented with the use of fragmentary pictures as test item stimuli, with the examinee able to increment the proportion of the picture presented.

Computer administration has been suggested as a means of measuring ability variables not convenient to test in paper-and-pencil format (Weiss, 1975). This will permit test designers and users to transcend the limits of traditional ability tests that measure verbal ability and logical, sequential analytical functions associated with the left hemisphere of the brain. Spatial perception, short-term memory, judgment, integration of complex stimuli, cognitive information-processing,

and other complex abilities may be measurable by exploiting the power and flexibility of the computer terminal as a testing medium. Cory (1978) has conducted exploratory research investigating computer administration of some novel item types. Valentine (1977) has discussed preliminary efforts directed toward computerized assessment of certain psychomotor abilities. Rimland and his associates (Lewis, Rimland, & Callaway, 1977) have used a computer to facilitate measurements of brain activity that may be related to ability variables. Rose (1978) is investigating measures of cognitive information processing skills using dynamic computer-administered problems as test items. All of the efforts just listed have shown some promise, but they must be considered as exploratory efforts that may or may not lead to developments that supplant or complement traditional methods of measuring psychological abilities.

REFERENCES

Baker, F. B., & Martin, T. J. FORTAP: A FORTRAN test analysis package. Educational and Psychological Measurement, 1969, 29, 159-164.

Bayroff, A. G., Thomas, J. A., & Anderson, A. A. Construction of an experimental sequential item test. ARI Research Memorandum 60-1, January 1960.

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests. ARI Technical Research Note 188, June 1967. (AD 655 263)

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models. RR 79-1. Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, February 1979.

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing. RR 74-4. Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, October 1974. (AD A001 230)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M., & Novick, M. R., Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-52.

Cliff, N. A basic test theory applicable to tailored testing. Technical Report No. 1. Los Angeles: Department of Psychology, University of Southern California, October 1975.

Cory, C. H. Interactive testing using novel item formats. In Weiss, D. J. (ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: Psychometric Methods Program, Department of Psychology, University of Minnesota, July 1978.

Jensema, C. J. An application of latent trait mental test theory to the Washington Pre-college Testing Battery. Doctoral thesis, University of Washington, 1972. (University Microfilms 72-20,871 Ann Arbor, Michigan.)

Jensema, C. J. An application of latent trait mental test theory. British Journal of Mathematical Statistical Psychology, 1974, 27, 29-48.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1977, 111-120.

Kalisch, S. J. The comparison of two tailored testing models and the effects of the models' variables on actual loss. Doctoral dissertation, Florida State University, 1974.

Krathwohl, D. R., & Huyser, R. J. The sequential item test (SIT). American Psychologist, 1956, 2, 419 (abstract).

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing. RR 74-3. Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, July 1974. (AD 783 553)

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. RR 75-1, Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, 1975.

Lewis, G. W., Rimland, B., & Callaway, E. Psychobiological correlates of aptitude among Navy recruits. NPRDC TN 77-7. San Diego: Navy Personnel Research and Development Center, February 1977.

Lord, F. M. A theory of test scores. Psychometric Monograph No. 7, 1952.

Lord, F. M. Some test theory for tailored testing. In W. Holtzman (ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (a)

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242. (b)

Lord, F. M. Robbins-Munro Procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (c)

Lord, F. M. Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce, & Suppes, Contemporary developments in mathematical psychology. San Francisco: W. H. Freeman, 1974. (a)

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264. (b)

Lord, F. M. A survey of equating methods based on item characteristic curve theory. RB-75-13. Princeton, N.J.: Educational Testing Service, 1975. (a)

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. RB-75-33. Princeton, N.J.: Educational Testing Service, 1975. (b)

Lord, F. M. A broad range tailored test of verbal ability. Applied Psychological Measurement, 1977, 95-100.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McBride, J. R. Scoring adaptive tests. In D. J. Weiss (ed.), Computerized adaptive trait measurement--problems and prospects. Minneapolis: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1975.

McBride, J. R. Research on adaptive testing 1973-1976: A review of the literature. Unpublished manuscript, May 1976. (a)

McBride, J. R. Simulation studies of adaptive testing: A comparative evaluation. Unpublished doctoral dissertation, University of Minnesota, 1976. (b)

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R. An adaptive test designed for paper-and-pencil administration. Paper presented at the 1978 convention of the Western Psychological Association, San Francisco, April 1978.

Olivier, P. A. An evaluation of the self-scoring flexilevel tailored testing model. Doctoral dissertation, Florida State University, 1974.

Owen, R. J. A Bayesian approach to tailored testing. RB-69-2. Princeton, N.J.: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogishe Institut, 1960.

Rose, A. M. An information processing approach to performance assessment; final report. AIR 58500-11/78-FR. Washington, D.C.: American Institutes for Research, November 1978.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph No. 17, 1969.

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 233-248.

Schmidt, F. L., & Gugel, J. F. The Urry item parameter estimation technique: How effective? Paper presented at the 1975 American Psychological Association Convention, Chicago, September 1975.

Seeley, L. C., Morton, M. A., & Anderson, A. A. Exploratory study of a sequential item test. ARI Technical Research Note 129, 1962.

Sympson, J. B. Estimation of latent trait status using adaptive testing procedures. Proceedings of the 18th Annual Conference of the Military Testing Association. Pensacola, Fla.: Naval Education and Training Program Development Center, 1976, 466-487.

Sympson, J. B. A model for testing with multidimensional items. Paper presented at Computerized Adaptive Testing '77, Minneapolis, July 1977.

Urry, V. W. A Monte-Carlo investigation of logistic mental test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: University of Washington, Bureau of Testing, 1971.

Urry, V. W. Computer-assisted testing: The calibration and evaluation of the verbal ability bank. Technical study 74-3, Research Section, Personnel Research and Development Center, U.S. Civil Service Commission, Washington, D.C., December 1974.

Urry, V. W. Tailored testing: A spectacular success for latent trait theory. Springfield, Va.: National Technical Information Service, 1977. (a)

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196. (b)

Vale, D. C. Problem: Strategies of branching through an item pool. In Weiss, D. J., (ed.), Computerized Adaptive Trait Measurement: Problems and Prospects. Research Report 75-5, Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, 1975. (AD A018675).

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing. RR 75-6. Psycometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, 1975.

Vale, D., & Weiss, D. J. A comparison of information functions of multiple-choice and free-response vocabulary items. RR 77-2, Psychometric Methods Program, Department of Psychology, University of Minnesota, April 1977.

Valentine, L. D. Adaptive test research. In Minutes of the Second Training and Personnel Technology Conference. Washington, D.C.: Office of the Director of Defense Research and Engineering, 29 March 1977.

Weiss, D. J. The stratified adaptive computerized ability test. RR 73-3, Psychometric Methods Program. Minneapolis: Department of Psychology, University of Minnesota, 1973. (AD 768 376).

Weiss, D. J. Strategies of adaptive ability measurement. RR 74-5, Psychometric Methods Program, Department of Psychology, Minneapolis: University of Minnesota, December 1974.

Weiss, D. J. Computerized adaptive ability measurement. Naval Research Reviews, November 1975, 1-18.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? RR 73-1, Psychometric Methods Program, Department of Psychology. Minneapolis: University of Minnesota, February 1973.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. RM-76-6. Princeton, N.J.: Educational Testing Service, 1976.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. D., & Douglas, G. A. Best test design and self-tailored testing. Research Memorandum No. 19, Statistical Laboratory, Department of Education, University of Chicago, June 1975.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

DISTRIBUTION

ARI Distribution List

4 OASD (M&RA)
2 HQDA (DAMI-CSZ)
1 HQDA (DAPE-PBR)
1 HQDA (DAMA-AR)
1 HQDA (DAPE-HRE-PO)
1 HQDA (SGRD-ID)
1 HQDA (DAMI-DOT-C)
1 HQDA (DAPC-PMZ-A)
1 HQDA (DACH-PPZ-A)
1 HQDA (DAPE-HRE)
1 HQDA (DAPE-MPO-C)
1 HQDA (DAPE-DWI)
1 HQDA (DAPE-HRL)
1 HQDA (DAPE-CPS)
1 HQDA (DAFD-MFA)
1 HQDA (DARD-ARS-P)
1 HQDA (DAPC-PAS-A)
1 HQDA (DUSA-OR)
1 HQDA (DAMO-RQR)
1 HQDA (DASG)
1 HQDA (DA10-PI)
1 Chief, Consult Div (DA-OTSG), Adelphi, MD
1 Mil Asst, Hum Res, ODDR&E, OAD (E&LS)
1 HQ USARAL, APO Seattle, ATTN: ARAGP-R
1 HQ First Army, ATTN: AFKA-OI-TI
2 HQ Fifth Army, Ft Sam Houston
1 Dir, Army Stf Studies Ofc, ATTN: OAVCSA (DSP)
1 Ofc Chief of Stf, Studies Ofc
1 DCSPER, ATTN: CPS/OCP
1 The Army Lib, Pentagon, ATTN: RSB Chief
1 The Army Lib, Pentagon, ATTN: ANRAL
1 Ofc, Asst Sect of the Army (R&D)
1 Tech Support Ofc, OJCS
1 USASA, Arlington, ATTN: IARD-T
1 USA Rsch Ofc, Durham, ATTN: Life Sciences Dir
2 USARIEM, Natick, ATTN: SGRD-UE-CA
1 USATTC, Ft Clayton, ATTN: STETC-MO-A
1 USAIMA, Ft Bragg, ATTN: ATSU-CTD-OM
1 USAIMA, Ft Bragg, ATTN: Marquat Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Tng Dir
1 USA Quartermaster Sch, Ft Lee, ATTN: ATSM-TE
1 Intelligence Material Dev Ofc, EWL, Ft Holabird
1 USA SE Signal Sch, Ft Gordon, ATTN: ATSO-EA
1 USA Chaplain Ctr & Sch, Ft Hamilton, ATTN: ATSC-TE-RD
1 USATSCH, Ft Eustis, ATTN: Educ Advisor
1 USA War College, Carlisle Barracks, ATTN: Lib
2 WRAIR, Neuropsychiatry Div
1 DLI, SDA, Monterey
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-MR
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-JF
1 USA Arctic Test Ctr, APO Seattle, ATTN: STEAC-PL-MI
1 USA Arctic Test Ctr, APO Seattle, ATTN: AMSTE-PL-TS
1 USA Armament Cmd, Redstone Arsenal, ATTN: ATSK-TEM
1 USA Armament Cmd, Rock Island, ATTN: AMSAR-TDC
1 FAA-NAFEC, Atlantic City, ATTN: Library
1 FAA-NAFEC, Atlantic City, ATTN: Human Engr Br
1 FAA Aeronautical Ctr, Oklahoma City, ATTN: AAC-44D
2 USA Fld Arty Sch, Ft Sill, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DI-E
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DT-TP
1 USA Armor Sch, Ft Knox, ATTN: ATSB-CD-AD
2 HQUSACDEC, Ft Ord, ATTN: Library
1 HQUSACDEC, Ft Ord, ATTN: ATEC-EX-E-Hum Factors
2 USAEEC, Ft Benjamin Harrison, ATTN: Library
1 USAPACDC, Ft Benjamin Harrison, ATTN: ATCP-HR
1 USA Comm-Elect Sch, Ft Monmouth, ATTN: ATSN-EA
1 USAEC, Ft Monmouth, ATTN: AMSEL-CT-HDP
1 USAEC, Ft Monmouth, ATTN: AMSEL-PA-P
1 USAEC, Ft Monmouth, ATTN: AMSEL-SI-CB
1 USAEC, Ft Monmouth, ATTN: C, Facl Dev Br
1 USA Materials Sys Anal Agcy, Aberdeen, ATTN: AMXSY-P
1 Edgewood Arsenal, Aberdeen, ATTN: SAREA-BL-H
1 USA Ord Ctr & Sch, Aberdeen, ATTN: ATSL-TEM-C
2 USA Hum Engr Lab, Aberdeen, ATTN: Library/Dir
1 USA Combat Arms Tng Bd, Ft Benning, ATTN: Ad Supervisor
1 USA Infantry Hum Rsch Unit, Ft Benning, ATTN: Chief
1 USA Infantry Bd, Ft Benning, ATTN: STEBC-TE-T
1 USASMA, Ft Bliss, ATTN: ATSS-LRC
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA-CTD-ME
1 USA Air Def Sch, Ft Bliss, ATTN: Tech Lib
1 USA Air Def Bd, Ft Bliss, ATTN: FILES
1 USA Air Def Bd, Ft Bliss, ATTN: STEBD-PO
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Lib
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: ATSW-SE-L
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Ed Advisor
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: DepCdr
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: CCS
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCASA
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACO-E
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACC-CI
1 USAECOM, Night Vision Lab, Ft Belvoir, ATTN: AMSEL-NV-SD
3 USA Computer Sys Cmd, Ft Belvoir, ATTN: Tech Library
1 USAMERDC, Ft Belvoir, ATTN: STSFB-DQ
1 USA Eng Sch, Ft Belvoir, ATTN: Library
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-TD-S
1 USA Topographic Lab, Ft Belvoir, ATTN: STINFO-Center
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-GSL
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATS-CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TE
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEX-GS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTS-OR
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-DT
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-CS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: DAS/SRD
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEM
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: Library
1 CDR, HQ Ft Huachuca, ATTN: Tech Ref Div
2 CDR, USA Electronic Prvg Grd, ATTN: STEEP-MT-S
1 HQ, TCATA, ATTN: Tech Library
1 HQ, TCATA, ATTN: AT CAT-OP-Q, Ft Hood
1 USA Recruiting Cmd, Ft Sheridan, ATTN: USARCPM-P
1 Senior Army Adv., USAFAGOD/TAC, Elgin AF Aux Fld No. 9
1 HQ, USARPAC, DCSPER, APO SF 96558, ATTN: GPPE-SE
1 Stimson Lib, Academy of Health Sciences, Ft Sam Houston
1 Marine Corps Inst., ATTN: Dean-MCI
1 HQ, USMC, Commandant, ATTN: Code MTMT
1 HQ, USMC, Commandant, ATTN: Code MPI-20-28
2 USCG Academy, New London, ATTN: Admission
2 USCG Academy, New London, ATTN: Library
1 USCG Training Ctr, NY, ATTN: CO
1 USCG Training Ctr, NY, ATTN: Educ Svc Ofc
1 USCG, Psychol Res Br, DC, ATTN: GP 1/62
1 HQ Mid-Range Br, MC Det, Quantico, ATTN: P&S Div

1 US Marine Corps Liaison Ofc, AMC, Alexandria, ATTN: AMCGS-F
1 USATRADOC, Ft Monroe, ATTN: ATRO-ED
6 USATRADOC, Ft Monroe, ATTN: ATPR-AD
1 USATRADOC, Ft Monroe, ATTN: ATTS-EA
1 USA Forces Cmd, Ft McPherson, ATTN: Library
2 USA Aviation Test Bd, Ft Rucker, ATTN: STEBG-PO
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Library
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Educ Advisor
1 USA Aviation Sch, Ft Rucker, ATTN: PO Drawer O
1 HQUSA Aviation Sys Cmd, St Louis, ATTN: AMSAV-ZDR
2 USA Aviation Sys Test Act., Edwards AFB, ATTN: SAVTE-T
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA TEM
1 USA Air Mobility Rsch & Dev Lab, Moffett Fld, ATTN: SAVDL-AS
1 USA Aviation Sch, Res Tng Mgt, Ft Rucker, ATTN: ATST-T-RTM
1 USA Aviation Sch, CO, Ft Rucker, ATTN: ATST-D-A
1 HQ, DARCOM, Alexandria, ATTN: AMXCO-TL
1 HQ, DARCOM, Alexandria, ATTN: CDR
1 US Military Academy, West Point, ATTN: Serials Unit
1 US Military Academy, West Point, ATTN: Ofc of Milt Ldrshp
1 US Military Academy, West Point, ATTN: MAOR
1 USA Standardization Gp, UK, FPO NY, ATTN: MASE-GC
1 Ofc of Naval Rsch, Arlington, ATTN: Code 452
3 Ofc of Naval Rsch, Arlington, ATTN: Code 458
1 Ofc of Naval Rsch, Arlington, ATTN: Code 450
1 Ofc of Naval Rsch, Arlington, ATTN: Code 441
1 Naval Aerospc Med Res Lab, Pensacola ATTN: Acous Sch Div
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L51
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L5
1 Chief of NavPers, ATTN: Pers-OR
1 NAVAIRSTA, Norfolk, ATTN: Safety Ctr
1 Nav Oceanographic, DC, ATTN: Code 6251, Charts & Tech
1 Center of Naval Anal, ATTN: Doc Ctr
1 NavAirSysCom, ATTN: AIR-5313C
1 Nav BuMed, ATTN: 713
1 NavHelicopterSubSqua 2, FPO SF 96601
1 AFHRL (FT) Williams AFB
1 AFHRL (TT) Lowry AFB
1 AFHRL (AS) WPAFB, OH
2 AFHRL (DOJZ) Brooks AFB
1 AFHRL (DOJN) Lackland AFB
1 HQUSAF (INYSD)
1 HQUSAF (DPXXA)
1 AFVTG (RD) Randolph AFB
3 AMRL (HE) WPAFB, OH
2 AF Inst of Tech, WPAFB, OH, ATTN: ENE/SL
1 ATC (XPTD) Randolph AFB
1 USAF AeroMed Lib, Brooks AFB (SUL-4), ATTN: DOC SEC
1 AFOSR (NL), Arlington
1 AF Log Cmd, McClellan AFB, ATTN: ALC/DPCRB
1 Air Force Academy, CO, ATTN: Dept of Bel Scn
5 NavPers & Dev Ctr, San Diego
2 Navy Med Neuropsychiatric Rsch Unit, San Diego
1 Nav Electronic Lab, San Diego, ATTN: Res Lab
1 Nav TrngCen, San Diego, ATTN: Code 9000-Lib
1 NavPostGraSch, Monterey, ATTN: Code 55Aa
1 NavPostGraSch, Monterey, ATTN: Code 2124
1 NavTrngEquipCtr, Orlando, ATTN: Tech Lib
1 US Dept of Labor, DC, ATTN: Manpower Admin
1 US Dept of Justice, DC, ATTN: Drug Enforce Admin
1 Nat Bur of Standards, DC, ATTN: Computer Info Section
1 Nat Clearing House for MH-Info, Rockville
1 Denver Federal Ctr, Lakewood, ATTN: BLM
12 Defense Documentation Center
4 Dir Psych, Army Hq, Russell Ofcs, Canberra
1 Scientific Advsr, Mil Bd, Army Hq, Russell Ofcs, Canberra
1 Mil and Air Attache, Austrian Embassy
1 Centre de Recherche Des Facteurs, Humaine de la Defense Nationale, Brussels
2 Canadian Joint Staff Washington
1 C Air Staff, Roval Canadian AF, ATTN: Pers Std Anal Br
3 Chief, Canadian Def Rsch Staff, ATTN: C CRDS(W)
4 British Def Staff, British Embassy, Washington
1 Def & Civil Inst of Enviro Medicine, Canada
1 AIR CRESS, Kensington, ATTN: Info Sys Br
1 Militaerpsykologisk Tieneste, Copenhagen
1 Military Attache, French Embassy, ATTN: Doc Sec
1 Medecin Chef, C.E.R.P.A.-Arsenal, Toulon/Naval, France
1 Prin Scientific Off, Appl Hum Engr Rsch Div, Ministry of Defense, New Delhi
1 Pers Rsch Ofc Library, AKA, Israel Defense Forces
1 Ministeris van Defensie, DOOP KL Afd Sociaal Psychologische Zaken, The Hague, Netherlands